DOCUMENT RESUME

ED 299 319 TM 012 324

AUTHOR Tatsuoka, Kikumi K.
TITLE Indices To Measure Stability of Rule Application.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE [88]
CONTRACT N00014-86-K-0290
NOTE 34p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Processes; Elementary Education; Equations (Mathematics); *Hypothesis Testing; Latent Trait Theory; *Learning Strategies; Mathematical Models; Misconceptions; *Psychometrics; *Test Items; Test Wiseness
IDENTIFIERS *Rule Application; *Slip Probabilities

ABSTRACT

When learning is taking place, students test their hypotheses and evaluate them, and modify their current theories on the basis of new information. This phenomenon is known as "hypothesis testing view" or "theory changes." Many students change their rules to another while they are taking a test. This study introduced a new concept called "slip probabilities" that quantifies to what extent a students' scores on an item deviates from the pure scores of his/her original rule. Algebraic relationships between this new slip-random variable and the traditional item score variable are discussed. Then, the indices to measure the stability of rule application and to compute the probability of having the perfect application of the rule are introduced and illustrated with a 64-item signed number test, which was given to 153 eighth graders. Thirty-nine erroneous rules of operations were represented by a set of response patterns on 32 items. A new random variable "u" that represents deviations from a pure rule by having a slip of 1 to 0 or 0 to 1 on an item helps explain the volatile cognitive activities that occur while a test is taken. Six data tables and seven graphs complement the text. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299319

Indices to Measure Stability of Rule Application

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KIKUMI TATSUOKA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

KIKUMI K. TATSUOKA

University of Illinois, Urbana-Champaign

Running head: Stability of Rule Application

TM012324

Acknowledgement

This study is partly supported by the Personnel and Training Research Program, Office of Naval Research under Contract No. N00014-86-K-0290.

Several of the analysis presented in this report were performed on the PLATO® system. The PLATO® system is a development of the University of Illinois, and PLATO® is a service mark of Control Data Corporation.

The author would like to express appreciation ot Bob Baillie for various computing programs and to Rose Nygaard for typing the manuscript.

Abstract

When learning is taking place, students test their hypotheses and evaluate them, and modify their current theories on the basis of new information. The phenomenon is famous as "hypothesis testing view" or "theory changes." Many students change their rules to another while they are taking a test. This study introduced a new concept called "slip probabilities" that quantifies to what extent a students' scores on an item deviates from the pure scores of his/her original rule. Algebraic relationships between this new slip-random variable and the traditional item score variable are discussed. Then, the indices to measure the stability of rule application and to compute the probability of having the perfect application of the rule are introduced and illustrated with a 64-item signed number test.

When a test is given as an integrated part of instruction, underlying cognitive processes governing students' performances on the test tend to be more volatile than those on standardized tests or ability tests. Glaser, Lesgold & Lajoie (1986) states that the students test their hypotheses and evaluate them, and modify their current theories on the basis of new information when learning is taking place. Many students change their erroneous rules (bugs) to another while they are learning (Tatsuoka, 1984a; Birenbaum & Tatsuoka, 1987). The first aim of this study is to introduce indices that measure stability of rule applications and the likelihood of having a perfect application of each rule. Then, these indices are applied to illustrate dynamic aspects of learning activities in the domain of signed numbers addition and subtraction operations.

Since the stability of rule applications is discussed in the context of slips deviated from the perfect application of a rule, a brief overview of Tatsuoka and Tatsuoka's Bug Distribution (1987a) will be given first. Then, stability index of rule R will be given as the expectation of the distribution function of slips around rule R (Rule R can be replaced by latent class R or rule class R). The probability of having the perfect application of rule R is given by the likelihood of zero pattern whose elements are all zeros. The algebraic relationship between the bug distribution for given R based on a slip variable u and the conditional distribution for a given θ based on item score x is given for explaining the fact that volatile cognitive activities can be observed in many students during the tests, and yet the test scores can be homogeneous and have high reliability values.

Brief Review of Bug Distribution

Tatsuoka and Tatsuoka (1987a) showed that slips deviated from a perfect application of rule R produces a cluster response around the pure response of R and it follows a compound binomial distribution with the probability of having a slip $p_{j|R}$ for item j (called a slip probability). Suppose u_j is a random variable such that $u_j=1$ if a slip occurs on item j and $u_j=0$ if not. If rule R is converted to a binary response pattern $R=(r_1 \ldots r_n)$ as stated in Tatsuoka, 1984b, then the slip probability of R on item j is given by Equation (1).

(1) $p_{j|R} = \text{Prob}\,(u_j = 1|R) = \text{Prob}\,(x_j \neq r_j|R)$

where x_j is observed score of item j and $u_j=|r_j-x_j|$.

Equation (1) can be rewritten as follows (Tatsuoka & Tatsuoka, 1987b):

(2) $p_{j|R} = r_j\ \text{Prob}\,(x_j = 0|R) + (1 - r_j)\ \text{Prob}\,(x_j = 1|R)$.

In Item Response Theory (IRT), the probability of having a score of one on item j for a given ability level θ is expressed by $P_j(\theta)$:

(3) $\text{Prob}\,(x_j = 1|\theta) = P_j(\theta)$.

Expressing the conditional probability given rule R in Equation (2) by a known quantity is not so simple. Tatsuoka (1985) showed that the numerator $f(\underset{\sim}{x})$ of IRT based caution index ζ introduced in Tatsuoka (1984) would be a convenient function because a response pattern $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)^1$ is expressed by the value of $f(\underset{\sim}{x})$ and the latent trait ability θ producing the response $\underset{\sim}{x}$. The mapping between $\underset{\sim}{x}$ and an ordered pair of $(\theta, f(\underset{\sim}{x}))$ is one-to-one for one parameter logistic model

and would be more so for other IRT models. Thus, the conditional probability given R, Prob $(x_j = 1|R)$ is rewritten by Equation (4).

(4) Prob $(x_j = 1|R)$ = Prob $(x_j = 1|\theta_R, f(\underset{\sim}{x}_R))$.

Equation (4) would be true for other person fit statistics such as the likelihood $L(\underset{\sim}{x}) = \prod_{j=1}^{n} P_j(\theta)^{x_j} Q_j(\theta)^{1-x_j}$, discussed by Drasgow, Levine and McLaughlin (1987), the expectation of $L(\underset{\sim}{x})$ introduced by Molenaar (1987) or Wright and Stone (1976) as long as these indices satisfy the condition described above in the context of $f(\underset{\sim}{x})$. If rule R is popular and used by many students, then the value of $f(\underset{\sim}{x}_R)$ would become near zero which is the expectation of $f(\underset{\sim}{x}_R)$ (Tatsuoka, 1985). Since the purpose of this study is to give a close examination on these typical misconceptions committed by testees in the population where the item parameters of IRT models were estimated, the values of $f(\underset{\sim}{x}_R)$ were expected to be near zero and thus the response patterns $\underset{\sim}{x}$ were assumed to be from the same population. Therefore the right side of Equation (4) can be rewritten by Equation (5).

(5) Prob $(x_j = 1|\theta_R, f(\underset{\sim}{x}_R))$ = Prob $(x_j = 1|\theta_R)$.

From both the Equations, (3) and (5), Equation (2) can be presented by Equation (6).

(6) $p_{j|R} = r_j$ Prob $(x_j = 0|\theta_R) + (1 - r_j)$ Prob $(x_j = 1|\theta_R)$.

It is well known that Prob $(X_j = 0|\theta_R)$ is IRT function $Q_j(\theta_R)$ and Prob $(X_j = 1|\theta_R)$ is $P_j(\theta_R)$. Therefore, the slip probability of rule R on item j, $p_{j|R}$ can be given in terms of IRT functions $P_j(\theta)$ as follows;

(7) $P_{j|R} = r_j\, Q_j(\theta_R) + (1 - r_j)\, P_j(\theta_R) = S_j(\theta_R)$.

For example, suppose rule R produces a response pattern of (111000) for 6 items, then the slip probabilities of 6 items are given as $Q_1(\theta_R)$, $Q_2(\theta_R)$, $Q_3(\theta_R)$, $P_4(\theta_R)$, $P_5(\theta_R)$ and $P_6(\theta_R)$ from Equation (7).

The responses that deviates from the perfect response or pure response resulting from the perfect application of the rule cluster around the pure response. The occurrence of these deviates (called slips earlier) across the items are assumed to be independent and hence these slips follow a probability distribution of the compound binomial form which is given by Equation (8).

(8) Prob (having up to κ slips from rule R) =

$$\sum_{s=0}^{\kappa} \left[\sum_{\Sigma u_j = s} \left\{ \prod_{j-1}^{n} S_j(\theta_R)^{u_j} (1-S_j(\theta_R))^{1-u_j} \right\} \right].$$

where $S_j(\theta_R)$ is given by (7), and s is the number of slips.

The relationship between the two binary random variables u_j and x_j can be expressed by (9) and (10) below;

(9) $x_j = r_j\,(1-u_j) + (1-r_j)\,u_j = |r_j - u_j|$

(10) $u_j = r_j\,(1-x_j) + (1-r_j)\,x_j = |r_j - x_j|$

by replacing u_j in Equation (8) by Equation (10), the probability distribution given by (8) can be rewritten with the variable x_j, score of item j as follows;

(11) Prob (having the total score of x_o)

$$= \sum_{x=0}^{x_o} \left[\sum_{\Sigma x_j = x} \left\{ \prod_{j=1}^{n} S_j(\theta_R)^{r_j(1-x_j)+(1-r_j)x_j} \, (1-S_j(\theta_R))^{1-r_j-x_j+2r_jx_j} \right\} \right].$$

Equation (11) is the cumulative distribution of total scores with the conditional probability of rule R, or latent class R. The difference between Equations (8) and (11) is whether the random variable u_j or x_j is used for describing the cumulative distribution of slips or scores, respectively.

Paulson (1985) recently formulated a latent class model in which each erroneous rule of operation was treated as a latent class. He further estimated the conditional probability of correct responses to each item given the respective classes. Paulson's model should produce very similar values of the conditional probabilities to those from the equation (11) if erroneous rules were yielded by typical sources of misconceptions.

Stability Index of Rule Application

One of the advantages formulating a bug distribution on a slip random variable (instead of scores) is that the measure of stability of rule application or the stability of a testee staying in a given same latent class is given by the expectation of slips of the probability distribution given in Equation (8). The expected mean and variance of the slips from rule R is given by Equation (12) and (13), respectively.

$$(12)\ \mu_R = \sum_{j=1}^{n} S_j(\theta_R) = \sum_{r_j=0} S_j(\theta_R) + \sum_{r_j=1} S_j(\theta_R) = \sum_{r_j=0} P_j(\theta_R) + \sum_{r_j=1} \theta_j(\theta_R)$$

$$(13)\ \sigma_R^2 = \sum_{j=1}^{n} S_j(\theta_R)\ (1-S_j(\theta_R)) = \sum_{j=1}^{n} P_j(\theta_R)\theta_j(\theta_R)$$

Tatsuoka and Tatsuoka (1987b) discussed two cases of expected number of slips, one is a slip of 0 to 1 and the other is 1 to 0.

The first term of the most right side of Equation (12) can be considered as the expected number of slips $0 \rightarrow 1$, from rule R, and the second term as the expected number of slips, $1 \rightarrow 0$. Equation (13) shows the variance of slips which is also the variance of total scores (Lord, 1980).

From Equation (9), the expected number of correct responses can be obtained and expressed by the sum of the number of ones in R, say r, and the difference between the expected numbers of the two kinds of slips, $0 \rightarrow 1$ slips and $1 \rightarrow 0$ slips as shown in Equation (14);

$$(14) \sum_{j=1}^{n} P_j(\theta_R) = r - \sum_{r_j=1} Q_j(\theta_R) + \sum_{r_j=0} P_j(\theta_R).$$

A Measure of Consistent Application of Rule R

If rule R is perfectly consistently applied, then the response pattern does not include slips and match with the pure pattern of rule R. As a result, the random variable u_j, $j=1,\ldots,n$ are all zeros and the probability of not having any slips is given by the $(n + 1)^{th}$ term of the compound binomial function as follows;

$$(15) \text{ Prob (having no slips)} = \prod_{j=1}^{n} (1-S_j(\theta_R))^{1-u_j} = \prod_{r_j=1} Q_j(\theta_R) \prod_{r_j=0} P_j(\theta_R).$$

Since the value of Equation (15) is extremely small, the geometric mean of (15) is taken in this study and it is defined as a quantity to measure the consistency of rule R. The range of the consistency of rule R is between 0 and 1 which are interpreted from lower consistency to a higher one.

$$(16)\ C_R = \left[\prod_{j=1}^{n} (1-S_j(\theta_R)) \right]^{\frac{1}{n}}.$$

In the next section,a newly defined stability index and the measure of consistent rule application will be illustrated with a test that is constructed for signed number addition and subtraction problems.

Method I

In order to examine the stability of rule application, a 64-item test comprised of two identical tests in which each test has 32 items of signed number problems was constructed. Thirty-two items were carefully constructed so that 39 erroneous rules of operation discovered in the previous studies (Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1981; ; Tatsuoka & Eddins, 1985) can be expressed by 39 different sets of response patterns uniquely. Thus, 39 erroneous rules of operations are represented by a set of response patterns on 32 items. Then the items were randomly ordered and placed in the first half of a 64-item test. The second half of the 64-item test was consisted of the same 32-items as those of the first half, but the items were randomly placed in positions independently from the first half of 32 items (called Subtest 1 and 2 hereafter). The same test was used in the study by Birenbaum and Tatsuoka (in press).

The 64-item test was given to the 153 eighth-graders at a local junior high school a year after they were taught signed numbers. The response patterns of 153 students were divided into two parts, the first half of 32 items and the second half of 32 items respectively. Two sets of the parameters of the two-parameter logistic model were estimated

separately from the first half and second half of the test, respectively.

Equating the Two Sets of Estimated Item Parameters

The two sets of estimated new parameters, $\left\{ \hat{a}_{1j}, \hat{b}_{1j} \right\}$ and $\left\{ \hat{a}_{2j}, \hat{b}_{2j} \right\}$ were first equated by minimizing the squared area of the two true scores of Subtests 1 and 2, $T_1(\theta)$ and $T_2(\theta)$ (Lord, 1980)

$$(17) \int_{-\infty}^{\infty} (T_1(\theta_1) - T_2(\theta_2)^2 d\theta_1 = \int_{-\infty}^{\infty} \left[T_1(\theta_1) - T_2(\alpha\theta_1 + \beta) \right]^2 d\theta_1$$

where $\theta_2 = \alpha\theta_1 + \beta$. The iterative computation for minimizing Equation (17) yielded almost identical true score curves of the two subtests, as shown in Figure 1. The values of $\alpha = .926$ and $\beta = -.080$ were obtained and the maximum distance between $T_1(\theta)$ and $T_2(\theta)$ was $\max_{\theta} | T_1(\theta) - T_2(\theta) | = 0.006$ at $\theta = -.700$. Table 1 summarizes the estimated item parameters and equated parameters of 32 items.

Figure 1 about here

Table 1

Analysis

One-hundred-fifty-three students were divided into four groups based on their ability levels $\hat{\theta}$ estimated from the 64 items. The ranges of θ values for the four groups were G1 (-1.730, -.756), G2 (-.731, .177), G3 (.187, .828) and G4 (.838, 1.816) where all zeros and all ones are deleted. The two kinds of statistics, $\hat{\theta}$ and IRT-based caution index ζ_2 (Tatsuoka, 1984b) were computed for each student in the four groups

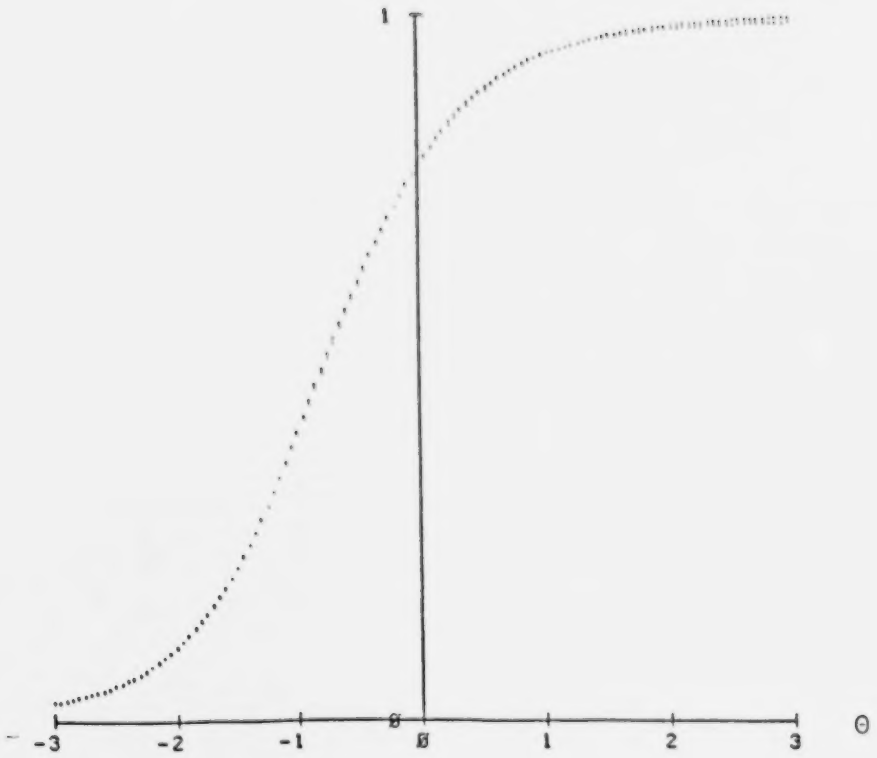

Figure 1

Figure 1. Equated true scores, $T(\Theta_1)$ and $T(\Theta_2)$ of Subtest 1 and 2.

Table 1: **Estimated Item Parameters Before and After Equated in the 64-Item Signed-Number Addition and Subtraction Test**

	1st Half, Equated		1st Half Before Equated to 2nd Half		2nd Half	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1)	.756	-1.544	1.159	-1.451	.699	-1.581
2)	1.633	- .974	2.697	- .995	1.512	- .966
3)	1.762	-1.075	1.419	-1.186	1.631	-1.074
4)	.848	- .041	.744	- .148	.785	.042
5)	1.442	- .343	1.097	- .599	1.335	- .283
6)	1.311	-1.243	1.392	- .961	1.213	-1.256
7)	1.353	- .903	1.202	- .832	1.253	- .889
8)	1.408	-1.317	.642	-1.838	1.303	-1.336
9)	1.250	- .504	.884	- .483	1.158	- .458
10)	1.697	- .340	1.525	- .359	1.571	- .313
11)	1.689	- .997	1.521	-1.072	1.563	- .990
12)	.728	-1.066	.921	- .943	.674	-1.065
13)	1.693	-1.111	1.087	-1.241	1.568	-1.113
14)	1.434	.483	1.476	- .135	1.328	- .435
15)	1.214	- .176	1.377	- .308	1.124	- .104
16)	1.076	- .185	.916	- .010	.996	- .113
17)	.728	-1.490	1.489	-1.314	.674	-1.523
18)	.969	-1.178	1.623	-1.124	.897	-1.186
19)	1.526	-1.291	1.280	-1.181	1.413	-1.308
20)	.851	- .256	.737	.020	.788	- .189
21)	1.137	- .063	.971	- .496	1.052	.018
22)	1.752	-1.134	2.245	- .973	1.622	-1.138
23)	.994	- .943	1.366	- .809	.920	- .932
24)	.922	-1.240	1.411	-1.087	.854	-1.253
25)	1.401	- .277	.851	- .624	1.303	- .213
26)	1.440	- .603	1.082	- .448	1.333	- .565
27)	1.285	-1.032	1.457	-1.016	1.189	-1.028
28)	1.353	- .716	.886	- .748	1.252	- .686
29)	1.159	-1.036	.859	-1.242	1.073	-1.032
30)	.908	- .462	1.398	- .480	.841	- .412
31)	1.031	- .515	1.186	- .203	.954	- .469
32)	.253	.998	.499	.151	.234	1.160

and then, Multivariate Analysis of Variance for two variables, $(\hat{\theta}_1, \zeta_1)$ and $(\hat{\theta}_2, \zeta_2)$, where $\hat{\theta}_1, \zeta_1$ and $\hat{\theta}_2, \zeta_2$ were obtained from the two 32-item subtests, respectively were carried out over the four groups. Table 2 summarizes the results.

Table 2

Both the variables θ and ζ accepted the nul hypotheses, $H_0 : \bar{\theta}_1 = \bar{\theta}_2, \bar{\zeta}_1 = \bar{\zeta}_2$ for each group except Group 3. Group 3 ($.187 \leq \theta < .828$) revealed that θ values were significantly different between the first half and the second half parts of the 64-item test but ζ values were not significantly different.

In the second analysis, grouping of subjects was done by the values of ζ, atypicality of response patterns obtained from the 64 items, as follows, G1 (-3.973, -1.326), G2 (-1.276, .111), G3 (.151, 1.063) and G4 (1.088, 7.463). The same MANOVA was carried out on (θ_1, ζ_1) and (θ_2, ζ_2). Table 3 shows the summary of the results and indicates that the

---

Table 3

---

members in Group 2 (-1.276, .111) have significant difference in θ values while stable in ζ values. Both the results of Tables 2 and 3 indicated the students whose ability levels were from the average to the level of slightly above and also the response patterns which were fairly typical to the sample showed the significant gain in θ values but the

Table 2: Summary Statistics of Multivariate Analysis of Variance on $(\theta_1\ \zeta_1)$ and $(\theta_2\ \zeta_2)$ When Four Groups are Determined by the Range of θ Values

Group	θ		ζ		Wilks' Lambda		
	F	P	F	P	Value	F	P
1	0.60	0.4420	0.02	0.8961	0.9900046	0.3483	0.7071
2	0.06	0.8092	0.10	0.7528	0.9979555	0.0686	0.9337
3	12.27	0.0008	1.30	0.2579	0.8303256	7.0500	0.0016
4	2.18	0.1445	0.46	0.5011	0.9627128	1.2975	0.2800

Table 3: Summary Statistics of Multivariate Analysis of Variance When Four Groups are Determined by the Range of ζ Values

Group	θ		ζ		Wilks' Lambda		
	F	P	F	P	Value	F	P
1	0.00	0.9992	0.00	0.9877	0.9999963	0.0001	0.9999
2	4.45	0.0383	0.86	0.3559	0.9429855	2.2673	0.1106
3	0.00	0.9876	0.69	0.4078	0.9905798	0.3471	0.7079
4	0.17	0.6850	0.11	0.7465	0.9974168	0.0945	0.9099

students in other categories showed no significant changes between the two 32-item subtests. Overall, the performances on the 64-item test were fairly stable in terms of scores obtained from the two 32-item tests. However it was puzzling why the students whose θ and ζ were in the range of $.187 \leq \theta < .828$ and $-1.276 \leq \zeta < .111$ performed significantly differently between the two subtests for θ values.

Investigation of Bug Stability or the Stability of Rule Application

The 39 response patterns corresponding to the description of bugs given in Tatsuoka & Tatsuoka (1981) are given in Table 4, and their centroids, expected number of slips, the total numbers of 1s and their consistency values are listed in Table 5.

Tables 4 & 5 about here

The range of θ varies from -4.31 to 2.84 while ζ values fall in between -4.44 to 7.58. The expected numbers of slips in both the cases of 1 to 0 and 0 to 1 vary from .47 to 16.93 out of 64 items. Rule 14 has, indeed, a very high expectation of slips and the lowest consistency among the bugs. This rule is very slippery and chances of observing the perfect application of rule 14 is very slim. In order to examine the relationship between θ and the expected number of slips, Figure 2, 3 and 4, 5 were obtained.

Figures 2, 3, 4, 5

Figure 2 is a graph of the number of expected slips $0 \rightarrow 1$ across various

Table 4: The Response Patterns Corresponding to the 39 Erroneous Rules (Bugs) Described in Appendix

Bug	Items																														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	1
2	1	1	1	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0	1	1
3	0	1	1	0	0	1	0	1	0	0	1	0	1	0	0	0	0	1	1	0	0	1	0	1	0	0	1	0	1	0	0	0
4	1	1	1	0	0	1	1	1	1	1	1	0	1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0	1	0	0	0
5	0	1	1	0	0	1	0	1	1	1	1	0	1	0	0	0	1	1	1	0	0	1	0	1	1	1	1	0	1	0	0	0
6	0	1	1	0	0	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	0	1	1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0
9	0	1	1	0	1	1	0	1	0	0	1	1	1	1	1	0	0	1	1	0	1	1	0	1	0	0	1	1	1	1	1	0
10	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0
11	1	1	1	0	0	1	0	1	0	0	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	0	1	1	1	0	0	0
12	1	1	1	0	0	1	0	1	0	0	1	1	1	1	0	0	1	1	1	0	0	1	0	1	0	0	1	1	1	1	0	0
13	1	1	1	0	0	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	1	1	1	0	0	1	1	1	1	0	0
14	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1
15	1	0	1	0	0	1	1	0	0	0	1	1	1	1	0	0	1	0	1	0	0	1	1	0	0	0	1	1	1	1	0	0
16	0	1	1	1	0	1	0	1	1	1	1	0	1	0	0	1	0	1	1	1	0	1	0	1	1	1	1	0	1	0	0	1
17	0	1	1	0	0	1	1	1	0	0	1	1	1	0	0	0	0	1	1	0	0	1	1	1	0	0	1	1	1	0	0	0
18	1	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0
19	0	0	1	0	0	1	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	0	0	0
20	1	0	1	0	0	1	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	1	0	1	0	0	0
21	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0
22	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0
23	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
24	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0
25	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
26	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0
27	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
28	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
29	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
30	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0
31	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
32	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
33	1	1	0	1	0	1	0	1	1	0	0	0	1	0	0	0	1	1	0	1	0	1	0	1	1	0	0	0	1	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	1	1	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	1	1	1	0	0	1	1	1	0	1	0	1	1
36	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
37	1	1	1	0	0	1	0	1	0	0	1	0	1	1	0	0	1	1	1	0	0	1	0	1	0	0	1	0	1	1	0	0
38	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
39.	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

Table 5: Summarizes the values of θ and ζ and $I(\theta)$ of the 39 rules, the expected numbers of slips 0 -> 1 and 1 -> 0 and the geometric mean of the probabilities of perfect execution of the rules (consistency of the rules) based on 64 items

			Expected Number of Slips			
Rules	θ	ζ	0 -> 1	1 -> 0	total 1s	consistency
1	2.84	.66	.47	0.00	64	.99
2	-.24	.21	12.76	11.69	48	.53
3	-.40	4.12	14.75	12.13	46	.51
4	-1.02	-2.83	12.53	15.68	24	.52
5	-.48	-3.06	13.63	15.32	40	.48
6	-.75	-1.02	14.07	16.77	32	.47
7	-.25	3.33	12.22	11.08	48	.54
8	.19	1.73	8.13	6.47	56	.66
9	-.48	.10	12.87	14.49	40	.50
10	-.10	.36	10.32	12.06	48	.52
11	-.80	-4.44	14.07	15.39	32	.49
12	-.64	-3.24	13.53	15.04	36	.50
13	-.46	-2.95	13.45	13.46	42	.51
14	-.96	7.58	16.93	15.79	30	.44
15	-.79	-.61	15.73	15.17	34	.47
16	-.58	1.02	14.37	15.38	38	.47
17	-.76	-2.70	14.53	15.00	34	.49
18	-1.10	-2.50	14.14	15.19	24	.49
19	-1.06	-.50	14.64	14.81	26	.49
20	-1.08	-2.49	14.27	15.88	24	.48
21	-1.36	.96	12.56	12.73	18	.52
22	-1.40	.46	12.42	11.82	18	.54
23	-1.88	1.29	8.36	7.16	10	.63
24	-1.41	1.12	12.56	11.71	18	.54
25	-1.46	-1.92	12.01	11.94	16	.53
26	-1.41	.39	12.59	11.67	18	.54
27	-1.41	.52	11.73	12.85	16	.54
28	-1.45	-1.87	11.82	12.08	16	.54
29	-1.45	3.61	11.84	12.10	16	.53
30	-1.59	1.16	10.84	10.23	14	.56
31	-1.63	.31	9.71	10.30	12	.58
32	-1.84	1.51	6.72	7.94	8	.66
33	-1.01	-.39	15.57	15.06	28	.47
34	-4.31	.45	0.00	.33	0	.99
35	-.63	5.68	15.88	15.65	38	.45
36	-1.75	1.40	6.59	9.26	8	.66
37	-.76	-3.10	13.56	16.00	32	.49
38	.17	-.92	7.86	7.87	54	.64
39	.56	.34	4.88	5.00	56	.73

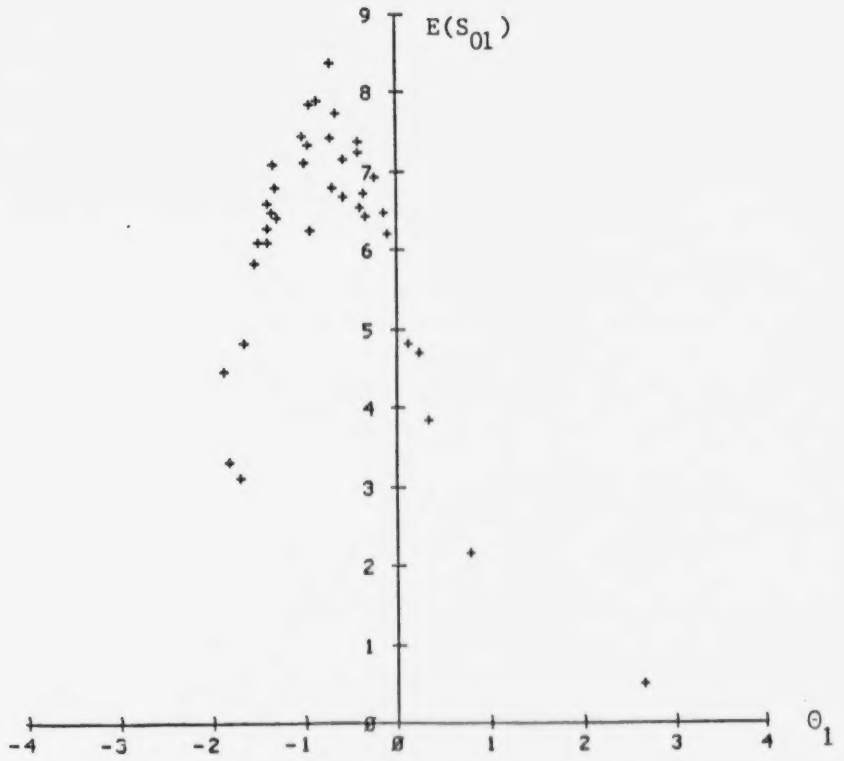

Figure 2

Figure 2. Expected number of slips 0 -> 1 of thirty-nine rules are plotted over ability levels θ_1 for Subtest 1.

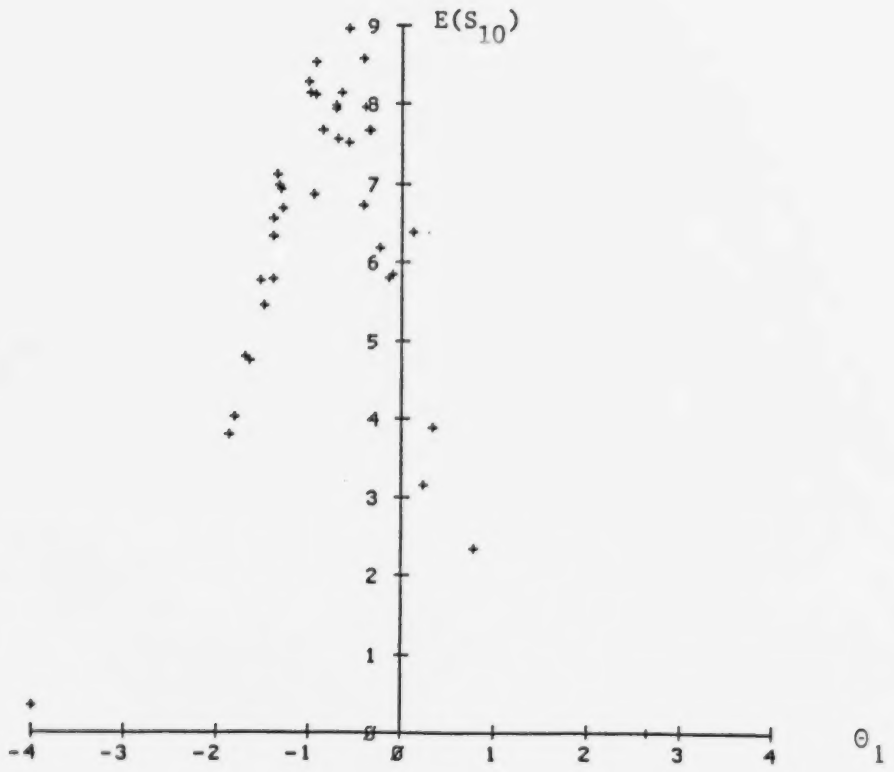

Figure 3

Figure 3. Expected number of slips 0 -> 1 of thirty-nine rules are plotted over ability levels Θ_2 for Subtest 2.

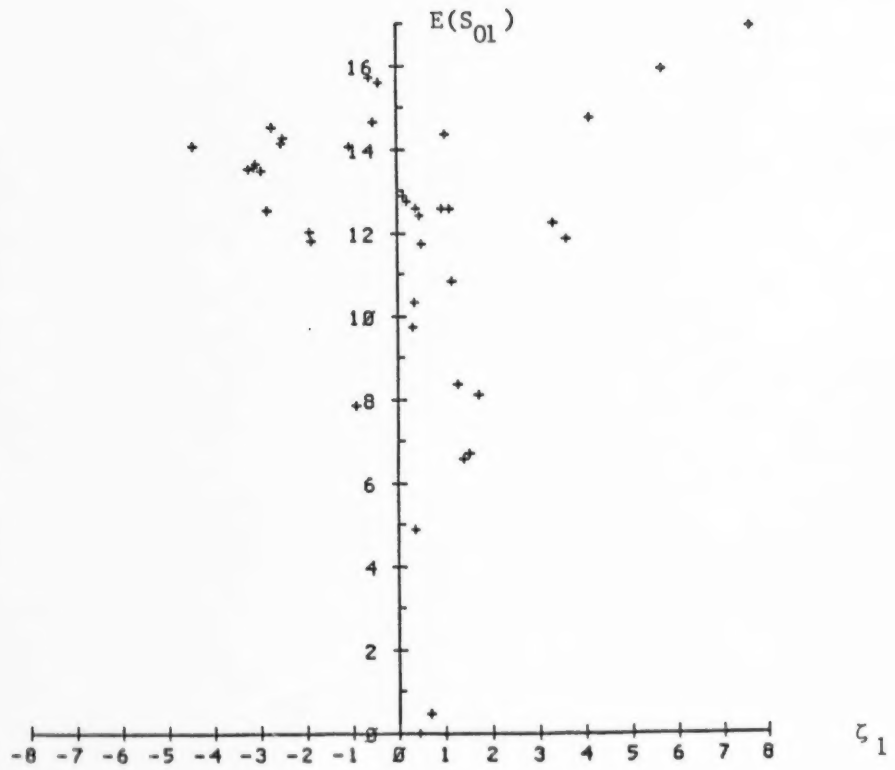

Figure 4

Figure 4. The expected number of slips 1 -> 0 of the thirty-nine rules are plotted over atypicality measure of ζ_1 for Subtest 1.

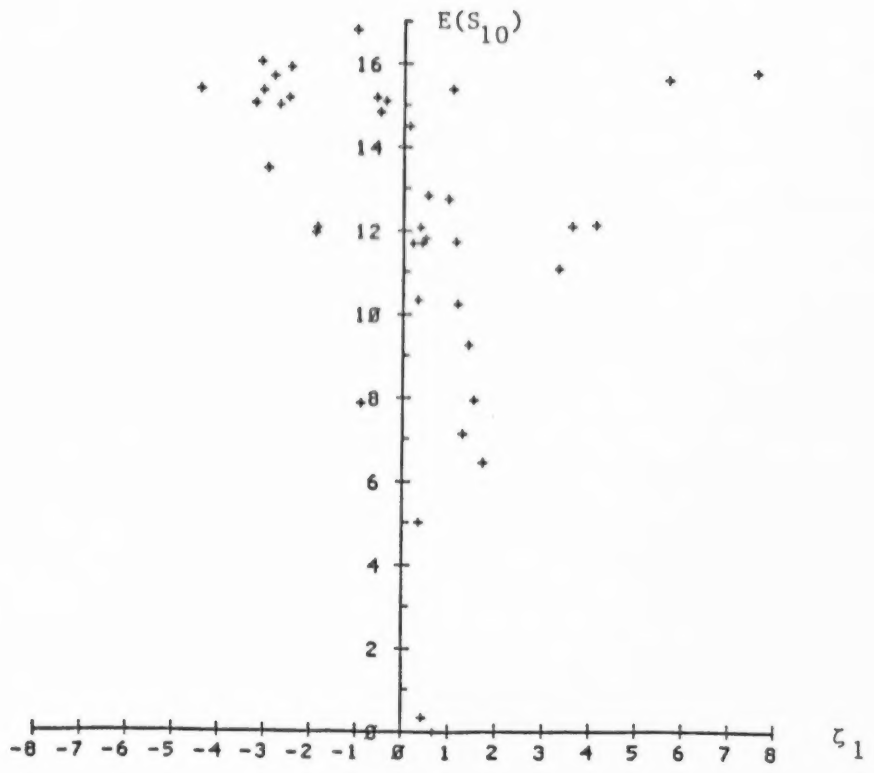

Figure 5

Figure 5. The expected number of slips 1 -> 0 of the thirty-nine rules are plotted over atypicality measure of ζ_2 for Subtest 2.

θ levels and Figure 3 shows the graph of the number of expected slips $1 \rightarrow 0$ across various θ levels for the first half of the test. Very similar plots are obtained from the second half of the test. Both the figures indicate that the rules occurring among high ability and low ability students have less numbers of expected slips while the rules often observed among average students have higher numbers of expected slips. That is, the rules committed by average students are unstable to some extent while those used by both the high and low ability students are fairly stable. This fact matches with the hypothesis testing view which is famous in the theories of cognitive psychology. Figures 4 and 5 show that the degree of atypicality of the rules does not influence the stability of rule application. However, the centroid of 39 rules

---

Figures 6 & 7

$\theta_{1\kappa}$ and $\theta_{2\kappa}$, $\zeta_{1\kappa}$ and $\zeta_{2\kappa}$ $\kappa = 1, \ldots, 39$, estimated from each half of 32 items correlate highly at .99 and .96, respectively. As can be seen in Figures 6 and 7, the squared error of mean differences of 39 rules are .037 for θ and .282 for ζ. The changes in θ values between the first and the second half of the 64-item test seem very small although the expected number of slips reach 8 or 9 in both the cases of $1 \rightarrow 0$ and $0 \rightarrow 1$.

Equation (14) in the previous section reveals that the true score $T(\theta_R)$ (Lord & Novick, 1968; Lord, 1980) at θ_R is the sum of the proportion correct of rule R, r/n and the difference between the expected numbers of slips $0 \rightarrow 1$ and $1 \rightarrow 0$ divided by the test length.

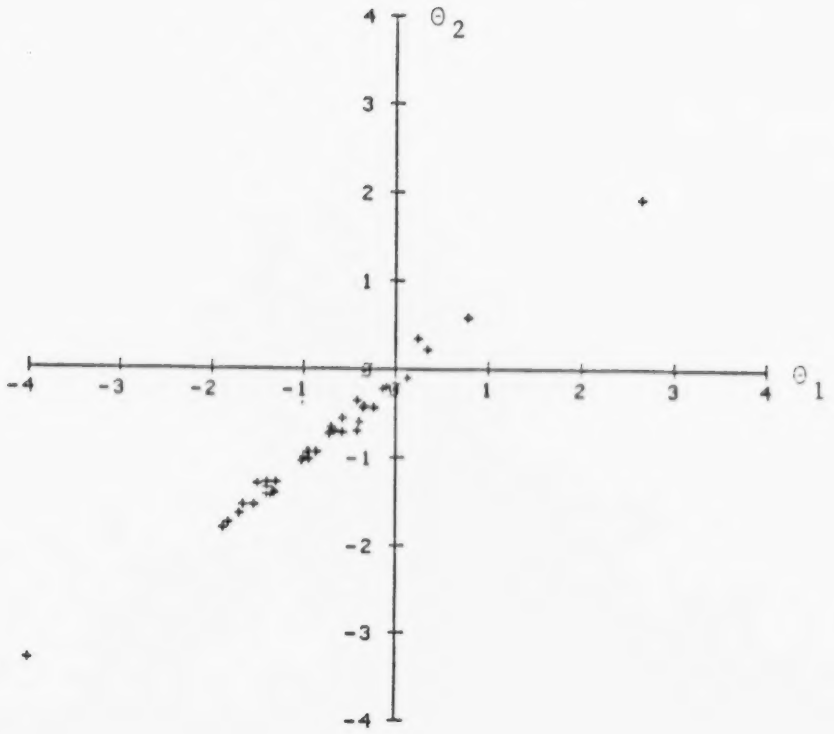

Figure 6

Figure 6. Plot of the thirty-nine rules with respect to Θ_1 and Θ_2.

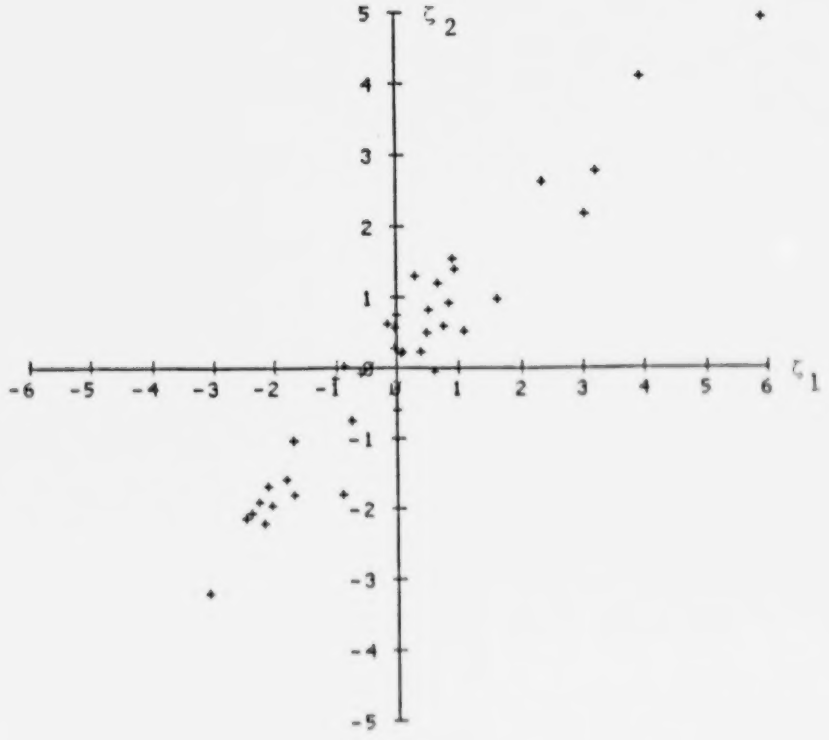

Figure 7

Figure 7. Plot of the thirty-nine rules with respect to ζ_1 and ζ_2.

This relation suggests that it is plausible to expect that students may have volatile cognitive activities and produce many slips in both the ways, 0 to 1 and 1 to 0, yet their true scores remain fairly invariant and stable. However, sample statistics based on item scores such as reliability of a test, means and standard deviations or correlations cannot reflect volatile cognitive activities, because the slips of $0 \rightarrow 1$ and $1 \rightarrow 0$ have a tendency of reserving the total number of ones unchanged and, thus, they have an effect of cancelling the scores increased by having a slip of 0 to 1 from those decreased by a slip of 1 to 0. As a result, the total scores remains fairly unchanged even if several slips are observed.

It was very difficult to treat the concept of "theory changes" in the traditional test theories. However, the introduction of a new concept, bug distribution in which slip variables and slip probabilities of rule R play an important role enable us to explain theory changes naturally, yet not contradicting with the traditional approaches. The above rationale will explain why the estimated θ values from the first and second half of the test remains almost unchanged although Figures 2 and 3 indicate fairly large numbers of the expected number of slips.

The puzzling result previously mentioned in Tables 2 and 3 is about high invariance of atypicality measure, ζ. Figure 3 also shows ζ_{1j} and ζ_{2j}, $j = 1, \ldots, K$ correlates high at .96. It is natural to assume that if the expected numbers of slips were large, then the values of IRT-based caution index would vary substantially between the first and

the second half of the test. As Tatsuoka (1984b) has shown, the numerator of ζ is given by $\sum_{j=1}^{n} \left[P_j(\theta_R) - x_j\right] \left[P_j(\theta_R) - T(\theta_R)\right]$. This quantity can be separated into two parts, the one which is independent from slips and the other which will be influenced by the number of slips.

$$(18) \quad -\sum_{j=1}^{n} \left[(r_j - u_j)(P_j(\theta_R) - T(\theta_R) \right]$$

If a slip would occur at item j whose absolute value of $P_j(\theta_R) - T(\theta_R)$ is very large, then the value of Equation (18) would change substantially. If a slip would occur at item j whose absolute value of $P_j(\theta_R) - T(\theta_R)$ is very close to zero, then the slip would not affect the magnitude of the value of Equation (18). Table 6 shows an example of an erroneous rule and its scores, slip probabilities, the values of $P_j(\theta) - T(\theta)$ and the rank order of $P_j(\theta) - T(\theta)$ from the largest to the smallest. Since the expected number of slips for this rule are 7.1 and 7.6 respectively, we assume that a student would have 7 slips of 0 to 1 and 7 slips of 1 to 0. Then, the change accounted for in Equation (18) would be -.592 and divide it by the denominator of ζ, variance of ζ given in Equation (19),

$$(19) \quad \left[\sum_{j=1}^{n} P_j(\theta_R)\, Q_j(\theta_R)\, (P_j(\theta_R) - T(\theta_R)^2 \right]^{\frac{1}{2}} .$$

Table 6

---

the actual change becomes very small, -.09. Indeed, it is negligible.

Table 6: The Summary Statistics of a Rule Selected from a List of 39 Rules

Items	Score	$S(\theta_R)$	$P_j(\theta)-T(\theta)$	Rank Order
13	0	.843	.233	4
11	0	.794	.184	7
1	0	.787	.177	8
17	0	.767	.157	10
27	0	.751	.141	12
29	0	.732	.122	14
4	1	.668	-.278	-3
32	1	.658	-.268	-4
16	1	.651	-.262	-5
25	1	.645	-.255	-6
10	1	.611	-.221	-7
20	1	.597	-.207	-9
9	1	.512	-.122	-12
31	0	.495	-.115	-13
30	0	.475	-.135	-11
14	0	.474	-.137	-10
26	1	.453	-.063	-14
28	1	.393	-.003	-15
5	0	.389	-.221	-7
12	1	.339	.051	17
23	1	.331	.059	16
15	0	.327	-.283	-2
7	1	.296	.094	15
21	0	.290	-.320	-1
18	1	.255	.135	13
24	1	.246	.144	11
2	1	.224	.166	9
6	1	.168	.221	6
3	1	.162	.228	5
22	1	.141	.249	3
8	1	.131	.259	2
19	1	.121	.269	1

The expected number of slips 0 -> 1 = 7.125

The expected number of slips 1 -> 0 = 7.600

Consistency, geometric mean of $\prod_{j=1}^{n} (1-S_j(\theta))$ = .486

$\hat{\theta}$ = -.5261, ζ = 2.150, the total score = 20

Therefore, we can safely conclude that ζ values across various θ-levels remain statistically unchanged, as can be seen in Tables 2 and 3 and Figure 7.

Conclusion and Discussion

Bug distributions introduced by Tatsuoka and Tatsuoka (1987a) played an important role for solving a conflict that has caused some alienation between cognitive psychologists and psychometricians. The traditional theories of psychometrics based on item scores was difficult to measure "theory changes" (Glaser, 1986), or to assess cognitive strategies used by students for solving a problem correctly. Linn (1985) discussed about serious technical barrier existing in the current measurement theory in Education and Psychology. This study showed that the current theories of psychometrics based on random variables representing scores of items or total scores, would be impossible to explain volatile aspects of cognitive activities underlying solving problems. But a new random variable u, that represents deviations from a pure rule by having a slip of 1 to 0 or 0 to 1 on an item enables us to explain the volatile cognitive activities while a test is taken. The relationships between the two variables x and u, between the bug distributions and the conditional distributions of rule R enable us to clarify why the test theories based on scores is impossible to measure "theory changes" that is a cognitive behavior expressed as "hypothesis testing" (Mayer, 1984; Birenbaum & Tatsuoka, 1987).

However, the estimation of slip probabilities given in this paper must be improved in future research.

References

Birenbaum, M., & Tatsuoka, K. K. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 20(1), 17-26.

Birenbaum, M., & Tatsuoka, K. K. (in press). On the stability of students' rules of operation for solving arithmetic problems. Journal of Educational Measurement.

Drasgow, F., Levine, M. V. & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal practical appropriateness indices. Applied Psychological Measurement, 11, 59-79.

Glaser, R. (1985). The integration of instruction and testing. Paper presented at the ETS Invitational Conference on the Redesign of Testing for the 21st Century. The Plaza, New York City.

Glaser, R., Lesgold, A., Lajoie, S. (1986). Toward a cognitive theory for the measurement of achievement (Technical Report).

Linn, R. L. (1985). Barriers to the new test designs. Paper presented at the ETS Invitational Conference on the Redesign of Testing for the 21st Century.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mayer, R. E. (1983). Thinking, Problem Solving, Cognition. W. H. Freeman and Company, New York.

Mayer, R. E. (1984). Thinking, problem solving, cognition. New York, NY: W. H. Freeman.

Molenaar, I. (1987). The many null distributions of person fit indices personal communication.

Paulson, J. A. (1985). Latent class representation of systematic patterns in test responses. (ONR Research Report) Portland, OR: Portland State University.

Tatsuoka, K. K. & Tatsuoka, M. M. (1981). Spotting incorrect rules in signed-number arithmetic by the individual consistency index. (Technical Report 81-4-ONR). Urbana, IL: University of Illinois, CERL.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4).

Tatsuoka, K. K. (1984a). Analysis of errors in fraction addition and subtraction problems (NIE Final Report). Urbana, IL: University of Illinois, CERL.

Tatsuoka, K. K. (1984b). Caution indices based on item response theory. Psychometrika, 49(1).

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Jour Educational Statistics, 12(1), 55-73.

Tatsuoka, K. K., & Eddins, J. M. (1985). Computer analysis of students' procedural "bugs" in an arithmetic domain. Journal of Computer-Based Instruction, 12(2), 34-38.

Tatsuoka, K. K., & Tatsuoka M. M. (1987a). Bug distribution and statistical pattern classification. Psychometrika, 52(2) 193-206.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987b). Item response theory, latent classes and rule space. (Research Report 1987-2-ONR, Final Report) Urbana, IL: University of Illinois.

Wright, B. D. & Stone M. H. (1979). Best Test Design. Chicago, IL: Mesa Press.